一种利用用户学习树改进的协同过滤推荐方法

马 莉
(天津外国语大学教育技术与实验室管理中心 天津 300204)

摘要:【**目的**】利用学习树中知识点的属性和学习访问序列, 对知识点进行预测评分, 进而进行用户相似性聚类以实施协同过滤推荐, 改进传统在线学习推荐方法, 提高推荐质量。【**方法**】对用户所学知识点属性、知识点学习访问序列、学习频率、学习时间进行标准化处理构建学习树; 基于学习树, 对树中知识点进行预测评分; 基于预测评分和知识点属性、知识点学习序列分别利用 Pearson 相似性和余弦相似性进行用户相似性计算, 利用 K 均值聚类方法进行相似用户聚类, 进而利用协同过滤推荐方法进行在线学习推荐。【**结果**】通过 F-measure 指标进行实验评价, 结果表明该方法与传统在线学习协同过滤推荐方法相比, F-measure 指标超过奇异值分解协同过滤 8.22%, 超过平均分预测协同过滤 3.75%。【**局限**】仅基于某在线学习平台的 52 456 条学生的学习记录和日志进行建模和测试, 未在其他数据集上进一步检验。【**结论**】解决了依赖用户评分进行协同过滤推荐的缺陷, 同时考虑了用户兴趣迁移对推荐准确率的影响, 对在线学习冷启动与可扩展性问题的解决具有较好的指导意义。

关键词: 在线学习推荐 协同过滤 学习树 学习访问序列
分类号: TP301.6 G35

1 引 言

近年来, 在线学习已经相当普及, 如何在海量的学习资源中找到用户最需要的学习资料成为在线学习的最大困难, 在线学习推荐系统的出现能够更好地为在线学习者准确定位其最感兴趣的学习资源。过去 10 年中, TEL 社团对在线学习推荐进行了深入研究[1], 推荐系统就是其设计的有助于在线学习的技术之一, 该技术能够为使用者创造个性化的学习环境。

目前主流的推荐算法主要有基于内容的在线学习推荐、协同过滤(Collaborative Filtering, CF)在线学习推荐以及混合推荐算法等。其中基于内容的推荐主要为用户建立一个学习兴趣模型, 将用户兴趣模型与资源属性进行匹配, 将匹配度最高的资源推荐给学习用户; 协同过滤推荐依据用户的兴趣, 对兴趣相似用户进行聚类, 在具有相似兴趣的用户之间进行交叉推荐, 是目前应用最为广泛也是最为成熟的在线推荐系统。以上两种方法各有优缺点, 基于内容推荐只考虑用户已经学习过的资源兴趣, 而无法发现其将来可能的学习兴趣; 协同过滤会存在新注册学习用户无法推荐的冷启动问题, 以及学习用户不愿意留下学习记录和学习评论的稀疏性问题。混合推荐方法致力于吸取前两种推荐算法的优缺点, 但如何将两种推荐算法更好地融合是混合推荐算法目前最大的挑战。

本文在协同过滤推荐算法的基础上, 通过对资源的属性、访问先后序列、学习频率、学习时间进行标准化处理, 为每位用户构建学习树, 基于学习树进行用户相似性计算及聚类, 进而实现在线学习协同过滤推荐。与传统的基于用户评分的协同过滤推荐算法相比, 本文方法具有较好的推荐准确性, 同时考虑了用户学习兴趣迁移对推荐准确性的影响, 并较好地解决了协同过滤推荐系统所存在的稀疏性和冷启动问题。

通讯作者: 马莉, ORCID: 0000-0002-9726-8286, E-mail: mali8321@tjfsu.cedu.cn。

2　研究背景

目前国内外学者提出的在线学习推荐方法主要有基于内容的在线学习推荐、协同过滤在线学习推荐以及混合推荐算法等，对每个算法的推荐原理及优缺点进行分析，如下：

(1) 基于内容的在线学习推荐

Khribi等[2]将用户最近的导航历史和学习资源内容进行相似性匹配，并在线自动生成学习建议。Sharif等[3]设计一个推荐框架，通过将学习资源的关键字与用户的学习兴趣标签进行匹配以实施推荐，并赋于学习资源权重，将学习资源根据重要性等级进一步排序。

基于内容的推荐算法，由于只考虑到对学习资源与用户兴趣特征进行匹配，而没有考虑用户之间的相似性，这样导致只会推荐用户已经学习过并有兴趣的资源[4]，而对那些用户未接触过的学习资源无法进行推荐。为了避免基于内容学习推荐的弊端，研究人员又提出了新的个性化方案，如协同过滤推荐技术[5]。

(2) 协同过滤推荐

绝大部分在线学习推荐系统利用协同过滤推荐方法实施推荐，其是目前应用最为广泛的推荐方法，可以被分为三类。基于近邻的协同过滤，通过用户对学习内容评分数据发现用户或学习资源之间的相似性，并对用户未评价的学习资源进行预测推荐。基于模型的协同过滤，利用用户评分矩阵，通过建立模型预测用户的评分。基于人口统计学的协同过滤，利用人口统计学特征进行用户相似性计算，并在相似用户之间进行学习资源推荐[6]。

在线学习环境中，学习资源媒体表现形式比较多样化，包括文本、超文本、图像、录像、音频和幻灯片，导致很难对学习资源的相似性进行衡量。通常在在线学习环境中，将依据用户对学习资源的偏好实施推荐。协同过滤是目前比较流行的推荐技术，但其有两个很明显的缺陷[7]。首要缺陷为稀疏性问题，很多用户不愿意对学习资源进行评价，导致进行用户相似性计算的基础数据缺失，进而影响推荐精度，很多研究人员通过数据挖掘技术获取隐性的有价值的信息以对稀疏数据进行补充。其次是冷启动问题，很多刚上线的学习资源，因评价数据较少，即使学习资源很有价值也很难被推荐到。Aher等[7]通过将学习资源进行分类，根据相关运算规则以缓解稀疏性和冷启动问题。基于在多元空间中学习资源的属性，Salehi等[8]提出树形模型为用户进行兴趣建模。并在学者树形模型中采用新的相似度计算方法产生推荐。实验结果表明他们提出的方法有效缓解了冷启动和稀疏性问题。

(3) 混合推荐方法

为了克服基于内容推荐和协同过滤推荐的不足，绝大部分的学者尝试采用某种方法将这两种方式混合进行推荐。Ge等[9]提出一种将基于内容和协同过滤推荐相结合的推荐方法。有些学者尝试将基于内容的推荐结果再次输入协同过滤推荐系统，并采用协同过滤推荐技术进行二次筛选；也有些学者尝试将协同过滤推荐结果再次经过基于内容推荐筛选[10]。因基于内容推荐和协同过滤推荐的推荐思想不同，如何将二者有效结合是混合推荐的核心内容，目前虽然大量学者提出了混合方法，但推荐精度和推荐效率值得进一步提高。

(4) 资源建模与用户建模

为更好地表示资源特征及用户学习兴趣，通常为资源及用户建立模型，Wang等[11]提出通过资源所在分类的属性表示资源；Kim等[12]进一步提出通过资源的内容特征属性建立资源模型，并依据特征属性的重要性进行排序；还有学者提出可以依据资源中出现的关键词数量为资源建立模型，关键词出现频率越高则该关键词表示该资源的权重越大；用户建模方面，最早提出的方法是可以依据用户的个人社会属性信息建立用户模型，这种方法可能会涉及个人隐私；Jalali等[13]通过利用用户访问过的资源特征建立用户模型，可以较好地反映用户学习偏好，这种方法只能发现用户已经表现出的历史偏好，而无法挖掘其未来可能的学习偏好；Albadvi等[14]提出通过聚类方法挖掘与某个用户具有相似偏好的用户簇，以该用户簇的共同偏好为当前用户建模，该方法取得了不错的效果。

3　研究框架

通过用户对资源的访问记录、访问时间、访问时长、访问频率等基本信息进行用户建模，用户模型由用户对资源的预测评分、资源访问序列、用户偏好转移组成；通过资源基本信息进行资源建模，资源模型由资源属性信息和属性权重组成。进而通过用户模型、资源模型构建用户学习树，通过用户学习树进行用户相似性聚类，并进行资源推荐度计算，进而实施推荐。

本文研究框架如图 1 所示:

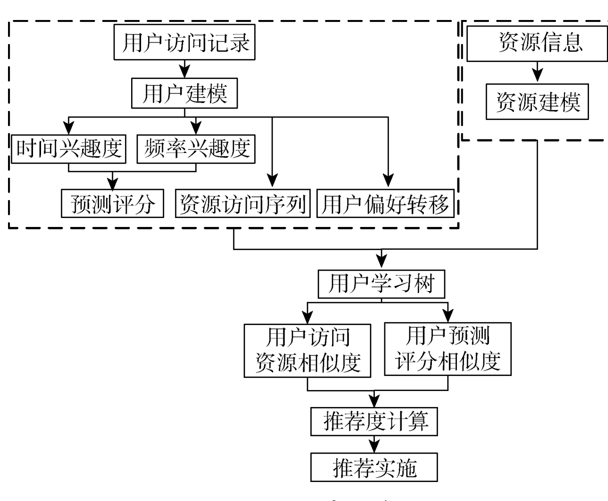


图 1　研究框架

4 推荐核心算法

4.1 资源建模

学习资源可以依据资源的类型进行分类, 如: 数学、物理、计算机科学等, 每一种类型可以进行进一步的细分, 如: 计算机科学又可以分为软件、计算机网络等, 可以依据学习资源所属的类型作为资源的属性。一种学习资源可以具有多种属性, 如: 某个学习资源既属于计算机科学, 也拥有数学知识属性, 同时还具有作者、学习类型等属性, 学习资源具有的属性应该是多维度的。用户可能是因为对资源的某个属性感兴趣而进行学习, 所以对用户因某个属性而进行资源访问的数量进行统计, 可以获得某个属性对该学习资源的权重贡献, 权重越大说明该属性相对于该资源更加重要, 该资源因该属性更加吸引用户。由此可以对资源进行建模如下: $M=[(Ak_1,Aw_1),(Ak_2,Aw_2)\cdots(Ak_m,Aw_m)]$, 其中 Ak_m 表示第 m 个属性的名称, Aw_m 表示第 m 个属性对该学习资源 M 的权重贡献, 本文设定: $Aw_1>Aw_2\cdots>Aw_m$, 同时 $\sum_{t=1}^{m}Aw_t=1$。如某个学习资源建模实例为: M=[(线性代数, 0.35), (概率论, 0.3), (硕士论文, 0.2), (某作者, 0.15)]。

4.2 用户建模

用户模型反映用户对于学习资源的偏好程度, 通常通过用户对资源的评分来反映。然而根据 Nielson[15] 的90-9-1理论: 90%的用户只是在网络上进行查找、阅读、浏览等, 而不原意参与互动(如对资源进行评价); 9%的用户可能偶尔参与网络互动, 但绝大部分时间他们只是在网络上浏览; 只有1%的用户在浏览的同时愿意参与网络的互动。因此, 很难得到较完整的用户评分对用户进行建模, 从而反映用户的兴趣。然而仍然可以通过提取用户对资源的访问记录, 进而处理得到用户对学习资源的兴趣偏好。本文主要将用户对学习资源的访问时间以及访问频率两个指标进行处理, 获取用户的学习偏好。

将用户访问过的学习资源序列建立集合 $s_{L_i}=\{M_1,M_2\cdots M_n\}$, 其中 L_i 表示用户 i, 集合中按照资源最近访问的时间顺序排序, 即 M_1 是最近访问过的资源, 即 M_n 是较久之前访问过的资源。通常来说用户对某个资源花费越多的时间, 表明该资源相对于用户越重要, 但有时候用户在资源上花费的时间多是因为该资源信息量大导致的, 用户在资源上花费的时间少也有可能是资源的信息含量少导致的, 综合考虑这些因素, 本文通过公式(1)对资源的用户访问时间进行处理, 得到用户对资源的时间兴趣度。

$$Time(L_i,M_j)=\frac{\dfrac{TotalTime(L_i,M_j)}{size(M_j)}}{\max(q\in s_{L_i})\left(\dfrac{TotalTime(L_i,M_q)}{size(M_q)}\right)} \quad (1)$$

$Time(L_i, M_j)$表示时间兴趣度, $TotalTime(L_i, M_j)$表示用户 i 在资源 j 上花费的学习时间, $size(M_j)$ 表示资源 j 的信息量, 通常是资源的存储容量。

用户对资源访问的次数越多同样可以表明该资源对用户有吸引力, 本文定义用户对资源的频率兴趣度如公式(2)所示:

$$Frequency(L_i,M_j)=\frac{Number_of_visits(L_i,M_j)}{\max(q\in s_{L_i})(Number_of_visits(L_i,M_q))} \quad (2)$$

其中, $Frequency(L_i,M_j)$ 表示频率兴趣度, $Number_of_visits(L_i,M_j)$ 表示用户 i 访问资源 j 的次数, $\max(q\in s_{L_i})(Number_of_visits(L_i,M_q))$ 表示访问次数最多的资源 M_q 的访问次数。

综合考虑某个学习资源的时间兴趣度和频率兴趣度, 利用公式(3)对其进行标准化处理, 得到用户对该资源的预测评分。

$$MR(L_i,M_j)=5\times Nor(Frequency(L_i,M_j)\times Time(L_i,M_j)) \quad (3)$$

其中, $MR(L_i,M_j)$ 为基于时间兴趣度和频率兴趣

度得到的用户i对资源j的预测评分，Nor()为标准化函数，将时间兴趣度和频率兴趣度处理为0-1之间的数值，此处预测评分为1-5分，随着用户对资源访问的变化，该预测评分也会随时更新。

需要指出的是用户对学习资源的最近访问时间能够反映出用户学习兴趣的动态转移，在E-Learning环境下，用户对学习资源的兴趣会动态变化，最近刚访问过的学习资源更能反映出未来用户的学习偏好，以往的用户学习模型对所有的学习资源同等对待处理，忽略了资源访问的时间顺序对用户偏好的影响。德国心理学家Ebbinghaus[16]提出的遗忘函数曲线反映了人类对新事物的遗忘规律，本文基于遗忘函数设计指数函数，反映用户对学习资源偏好的动态转移，如公式(4)所示：

$$h(x(M_j)) = \exp(-\lambda(x(M_j)-1)) \quad 0 \leqslant \lambda \leqslant 1, 0 < h(x) < 1 \qquad (4)$$

其中，$x(M_j)$表示用户 L_i 在其资源访问序列集合 s_{L_i} 中的次序，可以看出 $x(M_j)$ 在 s_{L_i} 中次序越靠后(值越大)，L_i 对 M_j 偏好越差，h(x) 也会越小。λ 为调节参数，反映用户对资源偏好的变化率，λ 越大 h(x) 变化越明显，即遗忘越明显。当 λ 取值为 0.95 时，h(x) 变化如图 2 所示：

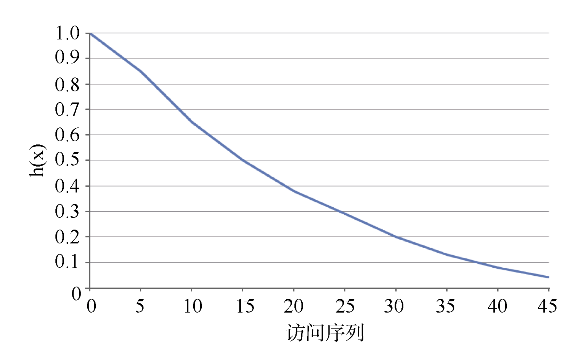


图 2　λ 取值为 0.95 时，h(x) 变化

基于用户对资源的访问序列，用户对资源的预测评分，反映用户偏好转移的 h(x)，本文为用户建立学习树模型如图3所示：

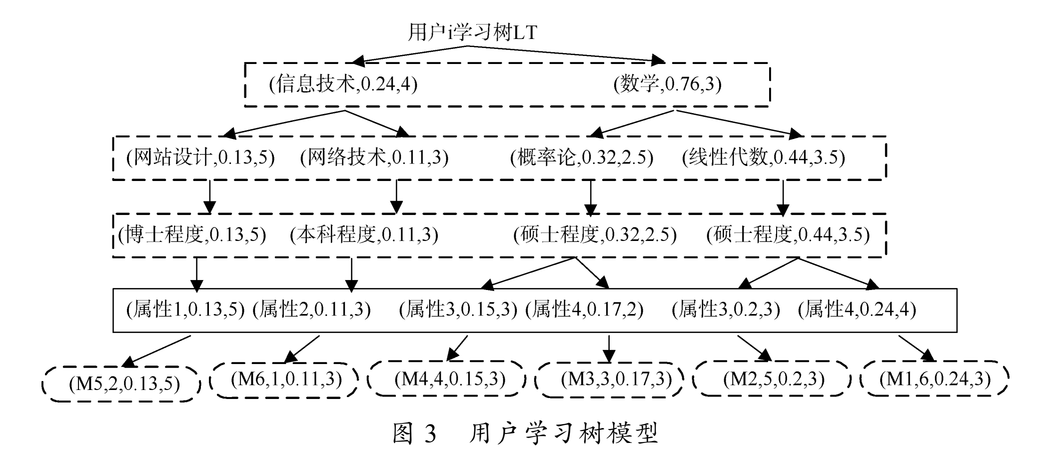


图 3　用户学习树模型

学习树是一个m+1层树状结构，m为用户访问过的资源属性的数量，树中最底层为叶子节点，表示用户访问过的一个学习资源，用一个四元组表示 $LT_{leaf} = \{MID, OR, NH, MR\}$，其中 MID 表示资源编号，OR 表示用户对该资源的访问次序，NH 表示用户访问过的资源的 h(x) 的标准化值，MR 表示基于时间兴趣度和频率兴趣度的预测评分。树中非叶子节点可以定义为一个三元组 $LT_{noleaf} = \{KA, NH, MR\}$，其中 KA 表示资源在该层的一个属性关键词，第i层节点的NH值可以表示为该节点第i+1层后继节点NH值的和，第i层节点的MR值可以表示为该节点子树中所有叶子节点MR的平均值。基于图3所示的学习树，用户对信息技术NH值为0.24，对数学的NH为0.76，反映出作者的兴趣发生了转移。

每当用户发生学习资源的访问，将会依据访问时长、频率和访问次序更新学习树，如学习树中无该学习资源节点，将会添加该节点并更新学习树。学习树动态更新过程如下：

(1) 用户访问了学习树上已存在的某资源节点M1(叶子节点)，记录该用户的学习时间和学习频率，同时修改该用户的资源访问序列。利用公式(1)可计算用户对该资源的时间兴趣度，利用公式(2)可计算用户对该资源的频率兴趣度，进而利用公式(3)可得到用户对资源的预测评分并更新叶子节点 $LT_{leaf} = \{MID,$

OR, NH, MR} 中的 MR；利用资源访问序列的变化可利用公式(4)重新计算动态兴趣度，并更新 NH，利用变化的资源访问序列更新 OR。因上层非叶子节点 $LT_{noleaf}=\{KA,NH,MR\}$ 中 NH 值和 MR 值为其下层节点的平均值，叶子节点更新完成后，学习树中的上层非叶子节点随之更新。

(2) 若用户访问了学习树中未存在的某资源节点(叶子节点)，该叶子节点 $LT_{leaf}=\{MID,OR,NH,MR\}$ 中的值计算过程同上，并依据资源属性在学习树中相应非叶子节点下新增叶子节点，并动态更新其上层非叶子节点。

4.3 相似学习用户聚类

目前绝大多数的协同过滤推荐系统多基于用户对学习资源的评分矩阵进行用户相似性计算并实施推荐。这种方法有两个缺点：

(1) 某些用户不愿意留下对资源的评分或者某些新上线资源还未有用户对其进行评价，这都会导致稀疏性问题。

(2) 此种方法过度依赖用户对学习资源的评分，而忽略了资源的属性以及用户学习的上下文，也会导致推荐精度降低。

用户学习树模型中包含用户学习资源属性、用户对学习资源的预测评分、学习资源的学习次序、用户学习偏好的偏移。本文提出基于用户学习树进行用户相似聚类的方法，该方法能够在不降低聚类效果的同时，有效避免传统协同过滤推荐的弊端。

本文提出基于学习树的用户相似性计算遵循以下三条原则：

(1) 学习树中学习资源的属性越相似，则用户的学习兴趣越相似。

(2) 学习树中用户对学习资源的学习顺序越相似，则用户的学习兴趣越相似。

(3) 学习树中用户对学习资源的预测评分越相似，则用户的学习兴趣越相似。

本文用户相似性计算分为两部分，基于学习树资源属性的相似性计算和基于学习树用户评分的相似性计算：

(1) 基于学习树资源属性的相似性计算 $sim_A(L_a,L_b)$ 如下：

$$sim_A(L_a,L_b)=\frac{\sum_{i\in AV(L_a,L_b)}MW_i\cdot NH_{ai}\cdot NH_{bi}}{\sqrt{\sum_{i\in LT(L_a)}MW_i\cdot NH_{ai}^2}\cdot\sqrt{\sum_{i\in LT(L_b)}MW_i\cdot NH_{bi}^2}} \quad (5)$$

其中，$AV(L_a,L_b)$ 表示用户 a 和用户 b 学习树中相同属性的交集集合，MW_i 表示学习树中第 i 层节点属性的权重，MW_i 所在节点层次越深该值越大，本文中定义 $MW_i=AW_i^{-1}$。NH_{ai} 表示用户 a 学习树第 i 层节点的 NH 值。

(2) 基于学习树资源预测评分的相似性计算 $sim_R(L_a,L_b)$ 如下：

$$sim_R(L_a,L_b)=\frac{\sum_{i\in L}|(MR_{ai}-\overline{MR_a})\cdot(MR_{bi}-\overline{MR_b})|}{\sqrt{\sum_{i\in L}(MR_{ai}-\overline{MR_a})^2}\cdot\sqrt{\sum_{i\in L}(MR_{bi}-\overline{MR_b})^2}} \quad (6)$$

其中，L 表示叶子节点集合，MR_{ai} 和 MR_{bi} 分别表示用户 a 和用户 b 对第 i 个叶子节点的预测评分，$\overline{MR_a}$ 和 $\overline{MR_b}$ 表示用户 a 和用户 b 的平均预测评分。

上述两个相似性计算公式只考虑了学习资源属性和预测评分，能够有效去除冷启动和稀疏性问题。

用户 a 和用户 b 的最终相似性如下：

$$LearnerSim(L_a,L_b)=\alpha\cdot sim_R(L_a,L_b)+(1-\alpha)\cdot sim_A(L_a,L_b) \quad (7)$$

其中，α 是 $sim_R(L_a,L_b)$ 和 $sim_A(L_a,L_b)$ 的权重，通过测试数据对 α 进行不同取值，发现 α 取值 0.7 时获得最好的推荐效果。

4.4 协同过滤推荐

推荐过程如下：

(1) 随着用户学习进程推进，为用户生成学习树并动态更新，学习树生成与更新过程见 4.2 节。

(2) 基于用户学习树中资源属性和预测评分进行用户相似性计算，用户相似性计算过程见 4.3 节。

(3) 对于某个用户 L_i 未接触过的资源 M_j 是否值得向该用户推荐，提出推荐度指标 $RD(L_i,M_j)$，推荐度指标计算如下：

$$RD(L_i,M_j)=\overline{MR_i}+\frac{\sum_{q\in LM_j}LearnerSim(L_i,L_q)\cdot(MR_{qj}-\overline{MR_q})}{\sum_{q\in LM_j}LearnerSim(L_i,L_q)} \quad (8)$$

其中，LM_j 为所有对资源 M_j 进行访问的用户集合，q 为该集合中某个学习用户，$LearnerSim(L_i,L_q)$ 为基于公式(7)得到的用户 i 和用户 q 的相似性，MR_{qj} 为用户 q 对资源 M_j 基于公式(3)得到的预测评分，$\overline{MR_q}$ 为用户 q 对所有资源预测评分的平均值。

基于推荐度指标，可以将推荐度最高的 Top-n 个用户的未学习资源推荐给学习用户。

5　实验评价

本文实验数据①为某国外在线学习资源的访问数据，该数据集包含完整的用户访问记录和资源基本信息，截取 2009 年 9 月–2011 年 2 月的访问数据。该段数据集包含 2 354 个用户的 52 456 条用户学习记录，包含 3 254 个学习资源，数据集包含基本信息完整。其中学习资源包含：资源编号、资源地址、上传时间、资源大小、适宜学习程度、资源分类、难易程度等基本属性，其中资源编号、资源分类可用于资源建模，用户访问日志包含：用户编号、访问路径、时间戳等信息，其中资源编号、资源大小、时间戳可用来计算时间兴趣度(公式(1))以及用户兴趣转移(公式(4))，用户编号、资源编号可用来计算频率兴趣度(公式(2))。评分日志包含：用户编号、资源编号、评分等基本信息。综合资源建模、时间兴趣度、频率兴趣度、用户评分等信息可用来用户建模(用户学习树)，基于用户学习树可进行用户相似性计算(公式(5)和公式(6))，并实施协同过滤推荐。

5.1　推荐精度、召回率和 F-measure

推荐质量的好坏通常用推荐精度和召回率两个指标进行测量，推荐精度为推荐的项目除以总推荐项目[17]。召回率为推荐的相关项目除以总相关项目(应当检索到的)。推荐精度和召回率的计算公式如下[18]：

$$\text{Precision}=\frac{|\{relevant_items\}\cap\{recommended_items\}|}{|\{recommended_items\}|}\tag{9}$$

$$\text{Recall}=\frac{|\{relevant_items\}\cap\{recommended_items\}|}{|\{relevant_items\}|}\tag{10}$$

由于召回率与精度是一对相互矛盾的指标，本文实验采用 F-measure 指标进行检验，这种方法混合了精度和召回率。

$$\text{F-measure}=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}\tag{11}$$

5.2　α取不同值对 F-measure 的影响

通过实验可以看到α取值 0.7 时，获得了较高的 F-measure 值，同时发现取值较小时推荐效果较差。

用户预测评分充分反映了用户对学习资源访问时长和访问频率的相似性，因而相比用户属性(用户访问资源相似)的相似性更能反应出用户学习偏好的相近程度：如图 4 所示，α 取值较小时用户属性相似性权重较大，F-measure 值偏低；当α 取值变大时用户预测评分权重变大，F-measure 值随之升高；但当α 超过 0.7 时，F-measure 值又有所降低。这是因为用户预测评分相似性只考虑用户对叶子节点(学习资源)的兴趣，而未从整个学习树角度考虑用户对资源分类的兴趣，所以α 取值过大反而对推荐质量有一定影响。

实验结果客观上反映了基于用户预测评分的用户相似性比基于资源属性的用户相似性对用户聚类效果影响要大。

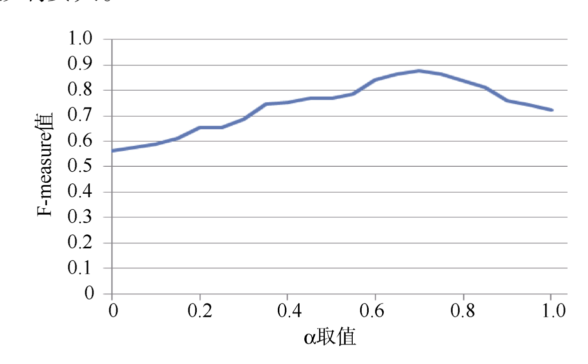


图 4　α不同取值对 F-measure 的影响

5.3　本文方法与其他推荐方法比较

目前消除稀疏性与冷启动问题较好的协同过滤推荐方法有 Sarwar 等[19]提出的经典奇异值分解协同过滤以及平均分预测协同过滤[17]两种推荐系统。

奇异值分解协同过滤推荐算法抽取原始用户评价矩阵最本质的特征，以提供一个简化的近似矩阵，这种方法消除了弱相关数据，从而降低了需计算数据的维度。由于推荐系统只对简化后的矩阵进行处理，只考虑降维后低维度数据，一定程度上降低了计算复杂度，是比较经典的协同过滤推荐算法之一。

平均分预测协同过滤推荐方法由 Devi 等[17]提出，对相似评分用户进行预聚类，基于聚类簇内用户的相似性对用户未评分数据进行预测，其本质为依据相似用户已评价产品评价值的相似性预测未评价产品评价值，该预测方法也取得了较好的效果，相比奇异值分解协同过滤在推荐精度上有一定提高，是目前推荐精度很高的主流协同过滤推荐方法。

①http://www.kdnuggets.com/datasets/index.html.

将本文方法与奇异值分解协同过滤以及平均分预测协同过滤进行比较，其 F-measure 值如图 5 所示：

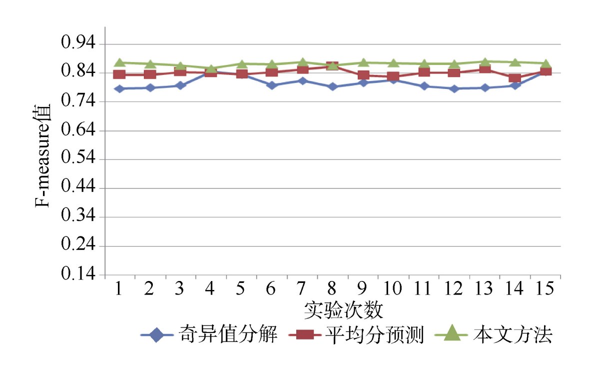


图 5　本文方法与奇异值分解、平均分预测的 F-measure 比较

实验结果分析如下：

(1) 本文方法的 F-measure 指标超过奇异值分解协同过滤 8.22%

(2) 本文方法的 F-measure 指标超过平均分预测协同过滤 3.75%

从实验结果来看，本文提出的推荐方法相比另外两种经典的推荐方法获得了较好的推荐质量，特别是相对奇异值分解协同过滤推荐算法效果好很多。奇异值分解推荐算法消除了弱相关数据，只抽取原始用户评价矩阵最本质的特征但其分解的效果对推荐质量影响很大，所以其推荐质量变化幅度较大，推荐效果不稳定。平均分预测协同过滤推荐效果要稍好一些，但其只能基于用户已评价的学习资源进行用户聚类并实施推荐，若用户未评价数据较多(稀疏性问题)，其推荐效果也会较差。删除部分已评价数据，提高稀疏性并进行实验，发现本文方法明显体现出优势，同时可以将“热点击”资源推荐给新注册用户，一定程度上消除了冷启动问题。实验结果如图 6 所示：

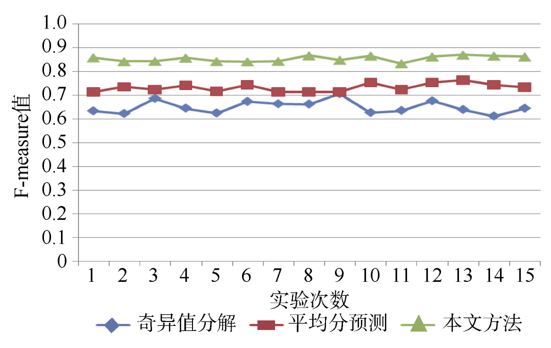


图 6　高稀疏性数据实验结果比较

6　结　语

本文提出的用户学习树充分考虑了用户学习资源的属性、用户学习资源的学习次序、用户对学习资源的预测评分及用户学习兴趣的转移，并基于此进行用户相似性聚类，通过学习用户的相似性进行学习资源推荐度计算，该方法能够有效地避免协同过滤推荐算法中的冷启动和稀疏性问题，实验评价结果表明本文提出的推荐方法在在线学习中具有较高的推荐质量。

参考文献：

[1] Verbert K, Manouselis N, Ochoa X, et al. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges[J]. IEEE Transactions on Learning Technologies, 2012, 5(4): 318-335.

[2] Khribi M K, Jemni M, Nasraoui O. Automatic Recommendations for E-learning Personalization Based on Web Usage Mining Techniques and Information Retrieval [J]. Educational Technology and Society, 2009, 12(4): 30-42.

[3] Sharif N, Afzal M T, Helic D. A Framework for Resource Recommendations for Learners Using Social Bookmarking [C]. In: Proceedings of the 8th International Conference on Computing and Networking Technology. IEEE, 2012: 71-76.

[4] Salehi M, Kamalabadi I N, Ghoushchi M B G. Personalized Recommendation of Learning Material Using Sequential Pattern Mining and Attribute Based Collaborative Filtering [J]. Education and Information Technologies, 2014, 19(4): 713-735.

[5] Salehi M, Kamalabadi I N. Hybrid Recommendation Approach for Learning Material Based on Sequential Pattern of the Accessed Material and the Learner's Preference Tree [J]. Knowledge-Based Systems, 2013, 48: 57-69.

[6] Chen W, Niu Z, Zhao X. A Hybrid Recommendation Algorithm Adapted in E-learning Environments [J]. 2014, 17(2): 271-284.

[7] Aher S B, Lobo L. Applicability of Data Mining Algorithms for Recommendation System in E-learning [C]. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics. 2012: 1034-1040.

[8] Salehi M, Kamalabadi I N, Attribute-based Recommender System for Learning Resource by Learner Preference Tree[C]. In: Proceedings of the 2nd International e-Conference on Computer and Knowledge Engineering. IEEE, 2012: 133-138.

[9] Ge L, Kong W, Luo J. Courseware Recommendation in

E-learning System [C]. In: Proceedings of the 5th International Conference on Advances in Web Based Learning. 2006: 10-24.

[10] Wan L, Zhao C. A Hybrid Learning Object Recommendation Algorithm in E-learning Context [J]. International Journal of Digital Content Technology and Its Applications, 2012, 6(18): 442-448.

[11] Wang S, Xie Y, Fang M. A Collaborative Filtering Recommendation Algorithm Based on Item and Cloud Model [J]. Wuhan University Journal of Natural Sciences, 2011, 16(1): 16-20.

[12] Kim K, Ahn H. A Recommender System Using GA K-means Clustering in an Online Shopping Market [J]. Expert Systems with Applications, 2008, 34(2): 1200-1209.

[13] Jalali M, Mustapha N, Sulaiman M N B, et al. OPWUMP: An Architecture for Online Predicting in WUM-Based Personalization System [A]. // Advances in Computer Science and Engineering [M]. Springer Berlin Heidelberg, 2009.

[14] Albadvi A, Shahbazi M. Integrating Rating-based Collaborative Filtering with Customer Lifetime Value: New Product Recommendation Technique [J]. Intelligent Data Analysis, 2010, 14(1): 143-155.

[15] Nielsen J. The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities [EB/OL]. [2015-09-01]. http://www.useit.com/alertbox/participation_inequality.html.

[16] Ebbinghaus H. Memory: A Contribution to Experimental Psychology [M]. New York: Dover, 1885.

[17] Devi M K K, Venkatesh P. Kernel Based Collaborative Recommender System for E-Purchasing [J]. Academy of Sciences, 2010, 35(5): 513-524.

[18] Klašnja-Milićević A, Vesin B, Ivanovic M, et al. E-learning Personalization Based on Hybrid Recommendation Strategy and Learning Style Identification [J]. Computers in Education, 2011, 56(3): 885-899.

[19] Sarwar B M, Karypis G, Konstan J A, et al. Application of Dimensionality Reduction in Recommender Systems [EB/OL]. [2015-10-25]. http://robotics.stanford.edu/users/ronnyk/WEBKDD2000/papers/sarwar.pdf.

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: mali8321@tjfsu.edu.cn。

[1] 马莉. RPD.re.txt. 学习资源属性原始数据.

[2] 马莉. ULD.log.txt. 用户访问日志信息原始数据.

[3] 马莉. URLD.rating.txt. 用户评分日志信息原始数据.

[4] 马莉. UD.u.txt. 用户信息原始数据.

[5] 马莉. TRT.doc. 测试集推荐结果–本文结果.

[6] 马莉. TRA.doc. 测试集推荐结果–平均分预测协同过滤.

[7] 马莉. TRS.doc. 测试集推荐结果–奇异值分解协同过滤.

收稿日期: 2015-11-09
收修改稿日期: 2016-01-20

Collaborative Filtering Recommendation Method Based on User Learning Tree

Ma Li
(Education Technology & Lab Management Center, Tianjin Foreign Studies University, Tianjin 300204, China)

Abstract: [**Objective**] This paper aims to improve traditionlal recommendation method and quality of E-Learning enviroment, which used attributes and access orders of resources in learning tree to predict learner's rate. The collaborative filtering recommendation was then carried out through similar learner clustring. [**Methods**] First, "attributes of resources""resource access order" "learning frequency and time" were standardized to construct users' learning tree and then predict resouces rating. Second, learner's similarity was calculated through Pearson and Cosine function respectivly based on predicted ratings. Third, K-means clustering method was used to group similar learners to establish collaborative filteing system for online E-learing. [**Results**] Compared with traditional collaborative filtering method, F-measure experimental result of the proposed method was 8.22% higher than the singular value decomposition CF and was 3.75% higher than the average score forecast CF. [**Limitations**] The proposed method was only tested on the dataset from one online learing platform with 52,456 students' learning records and access logs. More research is needed to examine the method with other data sets. [**Conclusions**] The proposed collaborative filtering recommendation system does not rely on learners' ratings and considers the influence of learners' interest changes. It could help us deal with the starting and expanding issues.

Keywords: E-Learning recommendation　Collaborative Filtering　Learning tree　Study access sequence

ProQuest SIPX 与 OpenStax、OpenSUNY 合作以促进"开放教育资源"获取

ProQuest一直致力于支持开放教育资源(OER)。通过此次合作, 使得OER的内容更容易被教职人员通过SIPX和Summon检索发现。现在, OpenSUNY OER的教材已被Summon索引, 其所有内容都将被索引, 并且还将与SIPX的课程相关联。随着这些开放资源更多地在校园学习管理系统中呈现, 教职人员将更容易选用这些资源。

通过此次合作, ProQuest、OpenStax和OpenSUNY能够为学生提供更多的选择, 以帮助他们减少获取课程材料资源方面的花费。"这次合作不仅能为高校提供更多高质量、免费的教材, 也使得这些内容更容易被发现。" ProQuest SIPX的总经理、联合创始人Franny Lee说道, "很高兴能够通过ProQuest增加相关课程资料, 我们一直为改善用户获取高质量、低花费的高等教育资源而不懈努力。"

OpenStax是一个非盈利的组织机构, 致力于帮助学生获取优质学习资源。"OpenStax致力于改善人们获取高等教育资源的现状。此次我们与ProQuest SIPX的合作能使得我们的内容资料被更多机构的读者获取到。" OpenStax的创始人Richard Baraniuk表示。

(编译自: http://www.proquest.com/about/news/2016/SIPX-Teams-with-OpenStax-and-OpenSUNY-to-Boost-Access.html)

(本刊讯)